AIR WAR COLLEGE

AIR UNIVERSITY

CUTTING THE CORD:

DISCRIMINATION AND COMMAND RESPONSIBILITY IN AUTONOMOUS LETHAL WEAPONS

by

Rob Trsek, Lt Col, USAF

A Research Report Submitted to the Faculty

In Partial Fulfillment of the Graduation Requirements

Advisor: Dr. Kimberly Hudson

13 February 2014

DISCLAIMER

The views expressed in this academic research paper are those of the author and do not reflect the official policy or position of the US government, the Department of Defense, or Air University. In accordance with Air Force Instruction 51-303, it is not copyrighted, but is the property of the United States government.

Biography

Lt Col Trsek is assigned to the Air War College, Air University, Maxwell AFB, AL. He is an Air Force senior pilot with 21 years of active service, formerly a command and control officer, AC-130H navigator, F-15C four-ship flight lead, F-4F instructor pilot, and Counter Land Operations project officer at the Air Force Doctrine Center. He is a graduate of The Ohio State University with a B.S. in Aircraft Systems, a graduate of Embry Riddle Aeronautical University with a M.S. in Aeronautical Science, a graduate of Air Command and Staff College with a M.A. in Military Operational Arts and Science, and a graduate of the School of Advanced Air & Space Studies (SAASS) with a M.A. in Air Power. He has served in the J5 directorate at USSOCOM and as the Operations Branch Chief at the Joint Deployable Analysis Team, Joint Staff J6, focusing on solutions to command and control and Combat ID in joint fires. He is also an adjunct professor to Embry Riddle Aeronautical University, teaching graduate and undergraduate aircraft accident investigation.

Abstract

The focus of this research is to examine the relationship between target discrimination (TD) and command responsibility (CR) as the primary barrier to the lawful use of autonomous lethal weapons under *jus in bello*. This paper begins with a thesis followed by three main points regarding the relationship and dependencies between TD and CR in the context of autonomous lethal weapons. Discrete roles in air-to-air and Air Interdiction are described that may permit autonomous systems to meet or exceed human thresholds in target discrimination and commensurate risk, followed by two brief case studies in fratricide to illustrate the main points submitted. Ultimately this research concludes that there is clear potential for autonomous lethal weapons to follow objective legal decision trees to meet rules of engagement criteria for the application of lethal force.

> As our understanding of the history of technology increases, it becomes clear that a new device merely opens a door; it does not compel one to enter.
>
> -- Lynn White Jr.

Introduction

Current force guidance documents indicate the US military intends to substantially increase its unmanned assets, seek mechanisms to reduce manning and spending under growing fiscal constraints, and counter anti-access/denial strategies employed by potential adversaries in order to maintain an advantage in armed conflict.[1] The confluence of these goals makes remotely piloted aircraft (RPA)[2] an attractive solution to cost-effective combat, simultaneously reducing the deployed footprint of forces, permitting extended duration sorties that translate into increased range or dwell time, and lowering the risk to (pilot) combatants.[3] As an extension of remote employment, autonomous unmanned air vehicles (UAV) would further reduce the number of human combatants, deny vulnerabilities in RPA command and control (C2) architecture, increase economies of scale through cooperative swarms, reduce the psychological stresses affecting RPA pilots by their displacement, and increase the speed of engagements.[4] A further advantage of autonomous weapons is their strict adherence to (literal) codes of conduct.

Autonomous lethal weapons (ALW) challenge standing moral and legal conventions, however, and their ultimate utility must rest upon the assurance that their employment will be possible within these bounds, or the conventions themselves (or accepted interpretation of them) must be challenged.[5] *Jus in bello* is the subset of just war theory focused on just conduct in war. It is the evolving product of global norms, cultures, and technology, which are reflected and codified in customary, national and international law, as well as conventions proper, such as Hague and Geneva. The focus of this paper is to examine the relationship between target

discrimination (TD) and command responsibility (CR) as the primary barrier to the lawful use of autonomous lethal weapons under *jus in bello*.

Thesis

In order to employ autonomous lethal weapons in discrete roles with assurance of target discrimination (TD) it is necessary to accept the foundation of command responsibility (CR) as observance of objective rules of engagement (ROE) rather than subjective operator judgment.

Target Discrimination and Command Responsibility

The Role of Command Responsibility with Respect to Failures

The first step in demonstrating the relationship between TD and CR is to begin with the very idea of CR, which is addresses failures in conduct, whether in discrimination, necessity or proportionality.[6] CR is the assurance and warning that commanders will be held accountable for failures to ensure lawful conduct of subordinates. It is important to note the idea of CR is deterrence, where failures may subject personnel to judicial proceedings, potentially criminal, with commensurate punishment for an offense. CR is also after-the-fact damage control, a means to show other commanders, combatants, or observant nations that unlawful conduct will be punished. CR does not fix the fault itself, resurrect those already killed, or mend the injured or property destroyed. CR does not alter the immediate consequence, it only acknowledges the failure to meet a standard and the continued ascription to that standard, whether sincere or with ulterior motives. CR then, removes and punishes the defective component (human) in the system, with the intent of reaffirming the standard and preventing future recurrence. Punishment does not remedy the consequence at hand.

The Utility of CR

The risk (R) assumed by employment of weapons is the product of the probability of a failure (P_f) in TD and the consequence (C) of that failure; $R = P_f \times C$. From this quantitative depiction, zero risk is only possibly by not employing weapons (man or machine), and therefore given war, some risk is always assumed absent perfect TD (zero P_f).[7] Accepting this deduction, it then follows that perfect TD would eliminate the utility of CR. That is, for a given target and associated consequence, if TD were perfect we would deny/degrade/destroy exactly what was intended and there would be no risk assumed, no danger of improper discrimination or the consequences of failure to properly discriminate. Absent risk there is no defective component to remove, the need to deter or punish is absolved. CR then is called for by the assurance that perfect TD cannot be obtained. More precisely, combatants are sure to make errors in TD and the enemy will try to amplify those errors, or consequences of failures to discriminate if it serves military, political or ideological purposes. CR is directly tied to levels of risk.

The Turing Test

In 1950 Alan Turing wrote *Computing Machinery and Intelligence,* which promoted his now famous "Turing Test." This test was to measure and distinguish between human and machine responses to identical stimuli, and it was the job of a third party human "witness" to determine which participant was man and which was machine. The basic idea being that if they were indistinguishable, the machine could be said to be intelligent.[8] In the same vein, in consideration of ALW and human ability to conduct TD we must consider their respective capabilities in the same manner as Turing – *performance* agnostic of man or machine. If ALW cannot discriminate as well as humans, their use should be seriously debated for roles where low confidence TD is shown. However, if ALW can discriminate with the same capacity or better

than their human counterparts, risk is unchanged or reduced and ALW should not be denied on grounds of TD alone.

> No doubt, it will be ethically very challenging to find acceptable roles and missions for military robots, especially for the more autonomous ones.
>
> -- Armin Krishnan

TD Assurance through Discrete Roles

Armin Krishnan, author of *Killer Robots*, suggests that it will be difficult to find high confidence TD roles for ALW, but does admit that technology holds the potential. More importantly, he admits that it is "absurd" to demand perfect TD accuracy given human failures in Western interventions, citing Kosovo, Afghanistan and Iraq.[9] Neither humans nor machines may be error free, but there are potential benefits to be gained through autonomy if machines can meet or exceed human performance in discrimination. Objections to ALW are grounded in 1) their inability to discriminate, or 2) moral objections to machines making lethal decisions. These are rational, logical responses with consideration of how these two ideas relate – clearly there is no interest in machines performing indiscriminate lethal actions.[10] Present technology may not allow a machine to distinguish a terrorist in urban garb from a civilian in his pajamas but certain roles and missions hold real promise for automation with high confidence of TD.[11]

Things, not People

If machines are not ready to discriminate between human combatants and noncombatants at the individual level, they have proven capable of discriminating between "things;" tanks, surface-to-air missile systems, fighter aircraft and so forth. The distinct missions of air-to-air combat and air-to-surface interdiction (AI) hold promise for ALW in that the environments are more sterile than the domain of human individuals, where targets are already engaged at long ranges in largely mechanized kill chains.

In air-to-air combat, fighters typically patrol a designated area searching for well-defined targets flying anticipated profiles with known (opposing) weapons and radars.[12] Enemy fighters can be engaged at ranges over 20 or 30 miles, well before the human pilot can visually identify the target.[13] Identification friend/foe (IFF) is accomplished via onboard and offboard means, to include radar signatures, electronic emissions, point of origin, flight profile, and other factors. A target is designated as hostile based on established ROE to include the above considerations, in addition to hostile act or intent. More explicitly, the pilot accepts the identification of "hostile" based on an electronic determination, onboard, offboard or a combination of the two, and engages a target with intent to kill long before it is ever seen. Indeed, successful engagement would mean opposing fighters never meet in the visual arena, and when they do, humans are subject to sensing errors as discussed below.

Accepting that close air support (CAS) missions with friendly and enemy combatants in close proximity may present difficult problems for autonomous TD today, AI offers a second potential role for ALW.[14] AI is typically conducted against pre-designated targets at fixed locations that have already met targeting criteria within intelligence and operational planning channels. This environment is much less dynamic than air combat; prior deliberate planning mitigates much of the risk and by definition has met the ROE. In this manner ALW appear no different than a cruise missile launched with authority to strike a designated target. What ALW can offer here is the ability to coordinate and cooperate with other ALW against target sets and emerging or dynamic threats to maximize survivability of the ALWs themselves and optimize weapons effectiveness against targets.[15] Clearly this is a more restricted role than suggested for air-to-air, however it emphasizes the point that the extension of CR from a cruise missile launch to an ALW with discrete targeting authority is a much smaller step than imagined. As Ronald

Arkin notes in his book on autonomous robots, "if a human being in the loop is the flashpoint of this debate, the real question is then, at what level is the human in the loop?"[16]

Moving CR from Subjective Judgment to Objective ROE

CR Theory and Practice

The overarching theory of command responsibility is that commanders and ultimately each combatant are responsible for their use of lethal force. While this theory is as valid for contemporary infantry as it was 4000 years ago, it simply does not translate to many modern weapon systems today, and unnecessarily restricts the use of ALW in discrete roles.[17] In practice, all combatants employ a *legal* decision tree based on laws of warfare in general and rules of engagement in particular.[18] As humans, each combatant may also apply their own subjective *moral* decision tree which may be in agreement with, or at odds with, their legal authority, and which has bearing on their decision to use force. However, complying with the laws of war and ROE are the legal foundation of using lethal force and by definition are the only requirements. It follows then, that if these rules governing the lethal use of force could be objectified for ALW, then ALW would be sufficiently equipped to meet the legal requirements.[19] In consideration of the risk equation and Turing test above, it is safe to conclude that if ALW can operate at a commensurate level of risk (pass the P_f Turing test), and follow objective ROE, then we must conclude that ALW can satisfactorily be used in lieu of human operators and reap the benefits that autonomy provides.

Human and Machine Error

It would be foolish to assert that either humans or machines are beyond error - both are replete with flaws that must be guarded against. While humans have always shown strength in flexibility and adaptation, our biology often fails us in repetitious, mundane or prolonged tasks

where machines excel. Machines, perhaps to a fault, do what they are designed to do with great precision. A complete examination of human factors is beyond the scope of this paper but to fairly compare humans and machines with respect to error and adherence to ROE a brief summary is required. Both human and machine errors primarily rest in sensation, cognition/ processing or execution. Because task execution is a consequent of cognition or machine processing (post-sensing), it is a dependent variable and not addressed here as a severable error. In fact, it is critical to understand that execution errors, to include decisions to use lethal force, can stem from sensory errors *or* processing errors *or* both.

Humans have long endeavored to extend our organic abilities through machines, so it isn't surprising that machines surpass us in many ways – they were designed to. The MQ-1 Predator ultra-wide field of view is 34°x 45° as compared to human 180°x 90° binocular vision, but our distant vision is poor by any standard of optics. Modern optical sights can see over many miles and may take advantage of different spectrums of light - most notably infrared or thermal.[20] As an example, the AIM-9X Sidewinder missile hosts an imaging infrared seeker that combines visual and IR spectrums for target ID and greater counter-countermeasure capability. The seeker uses an imaging database to identify the aircraft itself vice a prominent heat source, permitting autonomous ID of aircraft type.[21] Although machines have the upper hand in vision, particularly at range, even perfect sensing cannot guarantee appropriate processing.

Humans and machines process information in much the same way, but both suffer flaws inherent in their design. Both interpret their surroundings based on their prior programming; in machines this is explicit, deliberate, and hosts potential for standardization. Humans however, all have unique programming and are subject to biological and emotional flaws that can fail the lawful interpretation of ROE with consequent errors in task execution. Combatants may not

understand the ROE as written, they may not agree with it even if understood, they may desire to follow the ROE but simply fail to adhere to all measures (omission), or otherwise be influenced by stress, fatigue, fear, bigotry, racism, retribution, pity, empathy, depression or mental illness. To be sure, the horrors of war have shown the best and worst character in people, but it is impossible to quantify the lives saved or lost due to the play of human benevolence or malevolence in past wars. Machines, on the other hand, execute as programmed, subject to flexibility or flaws in software at the hands of their programmers. Both people and machines are subject to "reprogramming" however, where flaws in training or understanding of ROE can be adjusted when errors surface. Errors in interpretation of ROE with consequent behavior in machines could have widespread unfavorable effects however, making proper codification of ROE in programming critical to ALW.

Objective ROE

As previously stated, even perfect sensation cannot guarantee proper processing. The key to ALW following a legal decision tree rests on the ability to translate ROE into objective quantifiable rules to be followed autonomously. US ROE universally begin with a declaration that no rule can preclude the inherent right to self-defense, yet it is impossible for a machine to "feel threatened" in subjective terms. Hostile acts, hostile intent and so forth must be quantifiable if a machine is to adhere to such otherwise subjective rules, and for the discrete roles of air-to-air and air-to-surface AI suggested for ALW this is well within reach. In fact, the engagement factors that apply to the air-to-air environment and AI are readily transferable to machine language in the same terms that human pilots use and assess.[22] AI engagements would host similar factors regarding surface targets, but may also include statistical confidence levels of designated targets – say a tank or building – matching what was programmed in the database of

the ALW, or a circular error of precision (CEP) in terms of distance for a dynamic target.[23] As far as CR and risk are concerned, this latter point highlights that political leaders or military commanders could prescribe a certain risk level for ALW that cannot be accomplished uniformly across human combatants. Consider the AIM-9X seeker example as applied to an air-to-surface weapon. One could prescribe (program) a certain confidence level that a given target must match a database image, profile or location before weapons release were authorized, and in such a way that could vary by target type.[24] At risk of oversimplification, it could permit human authority at levels *above* the combatant/vehicle itself to determine exactly what level of risk was acceptable for a given mission. This would be done with the assurance that demands for high confidence TD and commensurate low risk of error could result in failure to accomplish the task. When an ALW can't meet the prescribed thresholds, it simply doesn't employ. Moreover, based on current automation's ability to account for multiple variables, consistency and reliability of data, we should expect machines to make better holistic decisions based on finite quantitative inputs than humans can, and in modern commercial aircraft such as the Airbus, they already do.[25]

Displacing CR

The considerations and assertions above are not to suggest that CR has no utility or is akin to dogma with no substantive need behind it. Humans should always retain control of the number, type and weight of variables that automation should consider and the thresholds required to legitimize lethal force. Demanding perfect discrimination may well result in mission failure. If greater risk is accepted to ensure employment (greater CEP or lower confidence in TD for example) then responsibility is *retained* at that decision level, it is not *lost* in the machine or in the software as some may assert.[26] In this way CR is displaced from the human pilot or ALW to the next higher authority. It is absurd to believe that a fighter pilot should be held liable for

errors in combat identification beyond visual range where his human senses are not in direct play. When an AWACS controller or offboard source labels a target as hostile based on its point of origin, altitude, airspeed, location, heading, lack of friendly IFF codes, emissions and/or other criteria, it should be with clear understanding that the pilot is releasing a weapon based on electronic information alone – information his/her senses cannot organically verify. The information available is measured against a legal threshold, and that quantified and objective information can be processed by a machine with greater accuracy and reliability than a pilot. Again, this is not to assert that mishaps will not occur. They will, as they have already occurred in other automated systems. As with human errors, mishaps should be investigated to determine the cause and proximate causes, whether they are sensory in nature or in processing, where faulty logic is subject to subsequent correction.

Case Studies / Illustrations

Blackhawk Fratricide

In April of 1994 two US Air Force F-15 fighters shot down two US Black Hawk helicopters in northern Iraq killing all 26 occupants. Findings from the official report concluded that procedures to include helicopter flights into daily operations were unclear and lacking, the AWACS mission crew commander was not current or qualified in accordance with Air Force regulations, the presence of the helicopters were not relayed to the F-15 pilots, IFF transponder settings were incorrect in the helicopters, and interrogation replies were intermittent and inconsistent.[27] Ultimately the F-15 pilots visually misidentified the Black Hawk helicopters as Soviet-made Hinds employed by the Iraqis, and shot them down.

It is impossible to prove an alternate history, but it is useful to examine the conditions and contributing factors that led to the Black Hawk shoot-down, and apply those findings to the

potential use of ALW in similar circumstances. This fratricide event is particularly useful as it occurred in a fairly sterile air-to-air environment in a typical employment mode for both the AWACS and fighters – one that could be expected for ALW in a possible future.

The AWACS crew was aware of the helicopters' presence in the area but failed to relay that information to the fighters. Non-standard terminology was used to identify the location and altitude of the radar contacts and AWACS never labeled the radar contacts as friendly or hostile. Without an offboard ID, the fighters were left to apply the ROE decision tree leading to potential engagement. While the explicit ROE is not contained in the report, at a minimum the following would be necessary: absence of friendly, presence of enemy (POE), a clear field of fire, flight operations within the no-fly zone. The helicopter pilots did not set the proper Mode 1 IFF code for the theater, but their transponders did reply to interrogations with the wrong code. Mode 4 IFF was successful once but subsequent interrogations by both fighters failed, for unknown reasons. For the pilots, absence of friendly was unclear where a single successful (albeit non-repeatable) Mode 4 reply would give most pilots pause. No pilot interview is given in the report but one could assume the visual ID pass was due to the inconsistent Mode 4 reply. During the visual pass, the lead pilot misidentified the helicopters as Hinds, simultaneously satisfying absence of friendly, POE, and flight operations in the no-fly zone. Adhering to the ROE, the pilots maneuvered for a clear field of fire and destroyed both helicopters.

Contributing variables on the part of the AWACS crew are manifold, however the lack of a hostile/friendly label displaced the burden of the ROE onto the fighters. This is not unusual, however the electronic information at the F-15s' disposal was ambiguous and the pilots were forced to rely on their own senses to satisfy the ROE. Erring in both sensation and processing,

the pilots misidentified the helicopters, failed to contact the Airborne Command Element (omission in ROE), with a consequent and tragic error in execution as a result.

In terms of ALW, we could posit the same circumstances, lacking an offboard ID the ALW would be forced to run its own decision tree based on sensation and processing. Again, one cannot presume the outcome would be any different; ALW sensors would be confronted with the same problem the humans met, namely that US Black Hawks with external fuel tanks look like Iraqi Hinds with external weapons pods. What would be different is that ALW sensors would be relying on superior optics and an imagery database, where the pilots were relying on biology and visual recognition training.[28] We could also expect ALW to be restricted to a specific TD confidence level prior to engagement, where failure to meet a threshold would terminate the engagement. ALW would also be unburdened by fatigue or a desire to log a "kill" as one might expect from human pilots.[29] If the ALW made the same sensing error as the human pilots, we could logically expect the same result. In this case, the ALW could perform *the same* as the human pilots, but there is significant potential it would have done *better* in sensing, and virtual certainty ALW would not fail the ROE by simple omission.

In terms of accountability, the lead pilot who misidentified the helicopters was granted immunity in order to testify, while the wingman was charged with 26 counts of negligent homicide; both pilots were removed from flying duty for three years but neither suffered more than administrative action and were permitted to resign and retire respectively.[30]

Iranian Flight 655

On July 3rd 1988 the USS Vincennes shot down Iranian flight 655, a civilian airliner, killing 290 passengers and crew. The circumstances surrounding the engagement by the AEGIS cruiser were complex; the escort of US oil tankers was born out of Iranian mining operations, the

USS Stark was recently attacked, Iranians had fired on US helicopters, and Iranian gunboats were harassing warships and merchants in the Strait of Hormuz. In the wake of the USS Stark incident US "commanders were given a revised set of ROE which clarified their authority to take positive protective measures when hostile intent was manifested. It was emphasized that they do not have to be shot at before responding and they have an unambiguous responsibility to protect their units and people." [31]

On July 2nd, USS Halsey had to "warn away" a threatening Iranian F-14 in the area. The following day, when Flight 655 took off from Bandar Abbas, a civilian *and military* use airfield, Vincennes was engaged with several small Iranian boats in the waters between Iran and Dubai, Flight 655's destination.[32] Flight 655 flew directly towards the Vincennes, and while it was within a civilian air corridor, it was late, more than 3 miles off centerline, and originated from the same airfield that launched Iranian F-4s against naval forces that April. To complicate matters, Vincennes could not identify any radar emissions, the contact bore a Mode 2 IFF code that indicated military *and* a Mode 3 IFF code that indicated civilian traffic, in addition to contradictory reports about aircraft ascent or descent.[33] The commanding officer, Captain Rogers, had less than four minutes to decide to engage the potential threat before it would be in range to threaten his own ship. The investigation concluded that, "In assessing what was reasonable performance under the circumstances it is imperative to have an emotional and intellectual feel for that picture [environment]." The Chairman of the Joint Chiefs of Staff remarked that "the Commanding Officer did what his nation expected of him in the defense of his ship and crew. This regrettable accident…was not the result of culpable conduct…"[34]

It would not be appropriate to substitute an ALW for the Vincennes and rewind the event to predict an outcome, as the roles and missions of UAVs and naval warships are not

interchangeable, nor could the outcome be certain. However, AEGIS cruisers are highly technological systems and can offer insight to human and machine errors within the scenario.

First, the identification of the threat was never definitively established, so like the Black Hawk fratricide, this scenario is confounded at the sensory level. The crew of the Vincennes had to rely entirely on electronic data with no opportunity to visually identify the target. The dual use airfield at Bandar Abbas and recent F-14 event with the USS Halsey obviated point of origin criteria for the purposes of identification, and set some precedent in the mind of the Vincennes commanding officer. "Flight 655 logically appeared to have a direct relationship to the ongoing surface engagement."[35] Airspeed and climb performance were consistent with both civilian and fighter aircraft. IFF was also ambiguous, showing both civilian Mode 3 and military Mode 2 responses, although the Mode 2 was never repeated following initial contact. Testimony shows diverse accounts regarding Flight 655 descent but AEGIS data recorded a consistent climb.

Application of a legal decision tree by ALW would rule out hostile ID by point of origin, airspeed, and altitude, but the flight profile would register as potentially threatening based on heading, aspect, range, closure and ambiguous ID as a potential F-14. What ALW may account for that the crew of Vincennes got wrong was the lack of targeting radar from the potential threat, and consistent information on the aircraft's ascent.[36] The holistic appraisal of this information may have yielded a more cautious approach, depending on the programmed TD accepted risk level. To be sure, an ALW would not target an ambiguous threat in self-defense unless the lowest threshold for TD were accepted, and in the mixed civilian/military environment over the Straits, it is unlikely the political cost would outweigh the loss of one or more ALW. The lesson here is not to suggest an ALW could replace the ship and loss of an ALW would be

preferable over fratricide in low confidence TD, rather that if an ALW were in such a position with ambiguous identification and unable to meet the ROE, it simply wouldn't engage.

The human side of the Vincennes story is more incriminating. Multiple officers from multiple ships in the theater testified to Captain Rogers' aggressive tendencies, his violation of ROE that prompted the gunboat response, his incursion into Iranian waters, and assertions that "his behavior was induced by a combination of physiological fatigue, combat operations, stress and tension." A more acute observation stated "the mind may reject incongruent data and facilitate misperception which promote internal consistency" [37] If these assertions have any merit, they point to the human failings of fatigue and stress, and cognitive bias in processing. Clearly the scenario was more than a sensory problem. It was a cognitive problem for the commanding officer, admitting that humans have a bias towards reaffirming their initial mental model of an event when assimilating new information. [31] Captain Rogers may have been reinforcing his model of an F-14 attack where a more agnostic approach to new information may have yielded a different conclusion. As with the Black Hawk incident, the ALW may have performed *the same* as Captain Rogers, but there is significant potential it would have done *better* through independent assessment of new information, absent an extant mental model, and without the influence of stress, fatigue, fear for his life and his crew, or bearing the responsibility for another USS Stark-like event. After killing 290 civilians, Captain Rogers was never charged with a crime, was awarded several medals, and was honorably discharged years later.[38]

Illustrations Conclusion

It is impossible to prove an alternate history but the above illustrations give some insight to human errors, and the potential for ALW sensors and processing to yield a different conclusion than their human counterparts. In these two scenarios, given the information present,

it is possible that ALW would not have engaged depending on prescribed levels of risk in TD. If nothing else is clear, we can be sure that humans have more limited sensory capabilities and are subject to cognitive bias, external influences, and failure to account for present *and absent* information in processing a mental model.

Conclusion

The Department of Defense has clear intentions to expand the use of unmanned systems in the military, and there are clear benefits to the use of autonomous systems in all domains. If we are to retain our ascription to standing legal conventions, codified in international law and military ROE, then it is critical to understand the relationship between CR and TD, and to determine precisely where the acceptable level of human responsibility belongs. This paper has demonstrated that the concept of CR doesn't need to be re-written to benefit from the use of autonomous systems in discrete roles, if the necessary legal code can be translated into objective decision trees for autonomous interpretation. For both AI and air-to-air engagements, the comparatively sterile environments lower the probability of TD errors and consequently the risk and utility of CR in those discrete roles. CR is already several steps removed from the operator in practice – it is naïve to believe that we are relying on biological sensing to fulfill ROE criteria, where the majority of information is electronically derived. CR need not be forfeited to acknowledge human failures in sensing and processing and to recognize and capitalize on machines where risk and TD permit it.

Appendix

Limitations

The scope of this research necessarily places limitations on the breadth and depth of tangential arguments and case studies contained herein. Space simply does not permit itemizing and deconstructing each principle of the commonly accepted laws of war, although discrimination, necessity and proportionality are all valid subjects for study in relation to ALW. Many would object to the use of an all-machine army against human adversaries as *prima facie* unethical or unlawful, and this extreme is worth critical examination.[39] Additionally, US cultural acceptance of or public confidence in ALW is not addressed directly. This is done with full recognition that cultural and technological developments are linked to one another, and our conduct in war and weapons in the US arsenal are subject to societal values/conditions. Most constrained are the case studies presented, which seek to draw out the main points of this paper, where each event culminated in volumes of data and interpretations not presented here.[40] Finally, history has shown that military efficiency or effectiveness have often displaced value-based considerations for conduct in war, where laws and conventions are set aside or derided in the name of military expediency.[41] While such expediency may be a primary driver of ALW, the focus here is on the extension of present concepts of TD and CR to ALW and not deliberate departures from or alternatives to existing legal conventions.

Definitions

Precision in language is vital to any meaningful discussion, particularly where philosophical ideas of morality and ethics are concerned, and even more so where technical terms and military jargon are applied. This paper focused on a narrow set of terms that are used

repeatedly and hopefully defy misinterpretation. Man-in-the-loop (MITL) and its cousins man-on-the-loop and man-out-of-the-loop (MOTL) refer to the presence and role of a human operator, if any, in the employment of a weapon system. The nuance between MITL and man-on-the-loop is that MITL *requires* a human for some portion of decision making, where man-on-the-loop is for oversight and intervention only *if needed*. ALW refers to, for the purposes of this paper, UAVs that have autonomy in lethal decision-making, as bounded only by their programming and operating MOTL, that is, without human oversight or intervention during mission execution.[42] Terms of ethics and morality are used generically and interchangeably, where no distinction between deontological, virtue ethics or consequentialism is intended. CR cannot be explicitly defined here, as just what that responsibility means is a component of this study. In general it refers to the body of customs, traditions and conventions that demand accountability for the lawful conduct of (self and) subordinates.[43] TD is the accurate distinction between lawful combatants and non-combatants, where non-combatants are accepted to be any person not meeting the definition of a combatant.[44] "High confidence" is used in reference to TD only to acknowledge that "low error" rates in discrimination are the focus of discussion, as failures to discriminate are assumed to be unacceptable from either humans or machines.

Notes

[1] United States. Department of Defense. Office of the Secretary of Defense. *Unmanned Aircraft Systems Roadmap, 2005-2030*. Washington, D.C.: Dept. of Defense Office of the Secretary of Defense, 2005, 56, 70.

[2] RPA pilots recoil at the use of the term Unmanned Air Vehicles (UAV) as it implies a pilot is not involved in the UAV mission. For the purposes of this paper, "RPA" will be used in reference to air vehicles with a human in the control loop (MITL) and UAV or autonomous UAV will be used to refer to air vehicles absent human control by design.

[3] Strawser, Bradley J. "Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles", *Journal of Military Ethics Vol. 9, No. 4*, (2010) 342-368. Strawser asserts that there is an ethical *obligation* to use UAVs as there is a "moral imperative to protect [the] agent if it possible [sic] to do so," and furthermore "I contend that in certain contexts UAV employment is not only ethically permissible, but is, in fact, ethically obligatory."

[4] Vick, Alan J., Lambeth, Benjamin S., *RAND: Aerospace Operations in Urban Environments* (2002), 201. In RAND's study, the authors conclude that "without a responsive and agile [human] command and control system, an elusive and adaptive adversary is likely to be there and gone before weapons can be brought to bear." Autonomous weapons that can navigate an objective legal decision tree to permit lethal use of force obviate the need for intricate chains of command that only serve the function of command responsibility at the expense of efficiency.

[5] Arkin, Ronald C., *Governing Lethal Behavior in Autonomous Robots* (CRC Press; 2009), 94-95. This paper asserts, and it seems that Arkin agrees, that ethical theory is "encoded in the Laws of War and Rules of Engagement." Both the ability to translate amorphous ethical rules into objective legal rules and the ability of autonomous machines to navigate the rule set are of vital importance. This paper is focused more on the latter as applied in discrete roles.

[6] Discrimination and CR is the focus of this brief paper, and the discrete roles suggested both pertain to environments with low (lower) risk of collateral damage. This is not to suggest that proper discrimination assures necessity or proportionality.

[7] Where $R= P_f \times C$, risk (R) is the product of two variables, the probability of failure (P_f) and the consequence (C) of that failure. Risk is low where the probability of failure is low *or* the consequence of failure is low. Risk is high where the probability of failure is high *or* the consequence of failure is high. Potential consequences will vary across multiple targets and not vary for a single given target, and therefore risk varies directly with the probability of failure to discriminate; a linear dependency.

[8] Turing, A. M., *Computing Machinery and Intelligence* (1950), 433-460.

[9] Dabringer, Gerhard, *Ethica Themen; Ethical and Legal Aspects of Unmanned Systems Interviews* (Publikation der Republik Osterreich, 2010), 55.

[10] Ibid., 45. Noel Sharkey, a computer scientist and professor at the University of Sheffield who chairs The International Committee for Robot Arms Control stated, "In a nutshell the ethical problem is that no autonomous robots or artificial intelligence systems have the necessary sensing and reasoning capabilities to discriminate between combatants and innocents."

[11] Vick, Alan J., Lambeth, Benjamin S., *RAND*, 201-2. The authors submit in their conclusion that in urban environments "it is *unlikely* that automated classifications of weapons, adversary personnel, or vehicles will be sufficiently reliable to permit lethal fires to be put automatically on targets." Emphasis original.

[12] Theater Operational Plans, or OPLANS are the foundation of theater deliberate planning and the textbook for theater combatants to become familiar with their area of responsibility and expected threats. Also referred to as an Enemy Order of Battle (EOB), mission-ready combatants are well versed in the hardware and tactics of their anticipated adversaries. It would be expected, not the exception, for a Pacific Air Forces pilot to be able to relay the full EOB for North Korean air forces, weapons loads, avionics capabilities, aircraft performance, and even the amount of training the adversary gets annually.

[13] Range offered as an unclassified, purposefully ambiguous, nominal range for a long-range air-to-air missile and supporting radar.

[14] Even CAS holds potential for ALW where pilots increasingly rely on electronic systems in the aircraft to aid in target separation from friendly forces. Combat Identification Server, or CID-Server is currently fielded in CENTCOM, which displays the nearest BLUFORCE tracker elements proximate to designated targets on the pilots Heads-Up-Display (HUD) and/or multi-function display (MFD). Consideration of and deconfliction with these known quantities/locations are within ALW capabilities.

[15] Alighanbari, Mehdi and How, Jonathan P., "Cooperative Task Assignment of Unmanned Aerial Vehicles in Adversarial Environments", *2005 American Control Conference* (2005). Galzi D. and Shtessel Y., "UAV Formations Control Using High Order Sliding Modes," *Proceedings of the 2006 American Control Conference* (2006). Note that "optimized" survivability among autonomous systems may include the necessary sacrifice of one UAV to ensure the survival of X number of others. In a human formation, deliberately sacrificing one fighter would be entirely unacceptable, forcing the formation into a position that could result in multiple losses. In this manner, a moral rule trumps logic and could lead to greater losses than rational, emotion-free decision making would permit.

[16] Arkin, Ronald C., *Governing Lethal Behavior*, 7.

[17] Ibid., 95. "Bounded morality ensures practicality, as it limits the scope of actions available and the situations in which it is permitted to act with lethal force." This paper, as well as the work of Matteo Turilli, asserts that those bounds can be translated from ROE to objective criteria. Turilli describes it as the translation of ethical principles into ethical requirements into ethical protocols, resulting in ethically constrained actors/agents. See also Turilli, M., "Ethical Protocols Design" *Ethics and Information Technology*, pp. 49-50, March 2007.

[18] Rules of engagement and special instructions (SPINS) are directive in nature and help combatants define in practical terms the preconditions and restrictions associated with the use of force, both in discretion and degree, with details specific to the local theater, adversary, noncombatants and operation at hand.

[19] This may sound like circular reasoning but it is not; the ability to translate ethical rules or protocols into a form that can be interpreted by a machine are a necessary antecedent to the machine executing within the bounds prescribed by those rules.

[20] Tidwell, M., "A Virtual Retinal Display For Augmenting Ambient Visual Environments", http://www.hitl.washington.edu/publications/tidwell/index.html (Human Interface Technology Laboratory). Nellis Air Force Base 57th Wing, "MQ-1 Predator Fast Facts," (September 2004). Smell, taste and touch are not addressed as they are not applicable to this topic, although hydraulics have proven far superior in terms of raw strength. Certainly applicable in combat but also not addressed directly is hearing, where machines again prove their superior abilities across audible frequencies and processing power to triangulate bearing and

distance – most useful in locating enemy snipers or mortar teams. See S. Moroz et al., "Airborne Deployment of and Recent Improvements to the Viper Counter Sniper System," Proceedings of the IRIS Specialty Group on Passive Sensors, Vol. 1, 1999, pp. 99–106, and L. S. Miller, "Counter Sniper Technology," Proceedings of the 5th Battlefield Acoustics Symposium, Ft. Meade, Md., September 23–25, 1997, pp. 681–692. RAND has also published a study.

[21] Jane's Defence Online, "Aim-9X," http://www4.janes.com/ (accessed January 2007).

[22] These factors include but are not limited to an unknown aircraft's heading, bearing, range, altitude, aspect angle, airspeed, vertical velocity, closure, IFF transponder codes, point of origin, location, radar type and mode (pulse repetition frequency (PRF), in search or target track) and so on. All are measurable, quantifiable, and objective.

[23] Where CEP is typically a measure of potential miss distance or error, it can also be used as a measure of distance from an expected location, useful in determining if a mobile target is likely to be the intended target based on its distance from where it was expected to be found.

[24] A machine could be programmed to only launch a "Maverick" AGM-65 at a potential target if it could determine from a desired sample set of X images that it were in fact a tank, or a T-72 tank, and so on. Alternate weapons could be prescribed for targets with known countermeasures in the most simplistic if->then format. Operations Research Systems Analysis (ORSA) specialists are accustomed to determining confidence levels for a determined measure of performance. For example one could require a 90% confidence that a system/objective ROE tree will yield 80% effectiveness or 20% discrimination failure. In this way, accepted risk can be strictly/explicitly controlled by political or military leaders as delegated. If this concept is foreign to the reader, consider a 100% confidence that a coin flip will result in 50% "tails."

[25] Trsek, Robert B., "Automation and Commercial Aircraft Safety" (2003), 8-9. Fifty two percent of all fatal crashes relate to poor pilot judgment. Research has shown that humans have several inherent faults when it comes to decision making, which are especially true in time compressed situations. First, humans tend to treat environmental cues with equal value. We are generally insensitive to cue *reliability* and therefore make poor predictions based on those false values. Taking things for face value as we do has led to the proposal that humans be used solely to identify *applicable variables*, and allow a computer or automated aid to derive the predicted outcome or course of action. Another area in which people fail is the bias we generate at the beginning of a sequence of events. Termed "cognitive tunnel vision," flight crews tend to adhere to the original mental model they create, despite subsequent evidence that would indicate the situation is otherwise. Humans will actually disregard new information if it doesn't fit their situational model. Once again, computers are better suited to this diagnostic task, as they use all available cues and have no bias towards that which occurs first. A last human failing worth mentioning is that we concentrate on *available* cues, rather than all cues both present *and* absent. What hasn't failed on an aircraft may be more of an indication of the problem than what has. Humans, however, will dwell on available cues and drive down the tunnel vision path, as we have limited resources to assess *all* aircraft systems in time constrained scenarios. See also Jensen, R., Aviation Psychology. USA: Gower Publishing Company (1989), 75-76.

[26] Arkin, Ronald C., *Governing Lethal Behavior*, 5. "If commanders are provided with the authority by some means to override the autonomous system's resistance to executing an order that it deems unethical, he or she in doing so would assume responsibility for the consequences of such action."

[27] United States. Department of Defense, "US Air Force Aircraft Accident Investigation Report: US Army UH-60 Black Hawk Helicopters 87-26000 & 88-26060 Executive Summary" (1994), 4-6.

[28] Ibid., 6. "Neither pilot had received recent, adequate visual recognition training."

[29] Government Accountability Office, "GAO/OSI-98-4 Review of USAF Investigation of Black Hawk Fratricide Incident" (1997), 8. The GAO report reviewed the official Air Force investigation and questioned the urgency of the F-15s to fire without additional identification passes, suggested the existence of an ongoing rivalry between the F-15 and F-16 communities, and cited the failure of the F-15s to inform the Airborne Command Element who had authority to terminate the engagement.

[30] Ibid., 11. If there were changes to the ROE following this event, the same could be expected following ALW error, denying the value of CR as a corrective mechanism resulting from human errors alone.

[31] United States. Department of Defense, "Investigation Report: Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988" (1989), 1.

[32] Ibid., 2.

[33] Ibid., 4-5.

[34] Ibid., 3, 7.

[35] Ibid., 2.

[36] Ibid., 4. USS Vincennes system data showed constant ascent throughout its flight, despite mixed testimony from the ship's crew, "decreasing in altitude, increasing in speed…"

[37] Ibid., Appendix.

[38] Wikipedia, "William C. Rogers," http://en.wikipedia.org/wiki/William_C._Rogers_III (accessed December 2013).

[39] The deterrent value of a "legal war" framework seems worth study as well. Restated, what would be the implications of autonomous systems that engage based on a legal construct with broadcast/forewarned assurance that moral impediments and human emotion would be absent at the combatant level. A purely legal war would almost certainly look a lot different than a legal war tempered by human involvement – restraint, empathy, ethical judgments, etc. A thesis might suggest belligerents would be less inclined to engage with known autonomous responses - people are subject to doubt and bluffing – machines are less forgiving.

[40] The Black Hawk Investigation is over 3,000 pages in length. The Government Accounting Office reports on the same are 60 pages (November 1997) and 8 pages (June 1998).

[41] Walzer, Michael. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. *4th ed.* (New York: Basic Books, 2006), pp. 231-232. Walzer asserts there are four ways that states deal with the tension between just conduct in war and just warfare when conventions or laws stand in the way of immediate interests: 1) The convention is simply set aside (derided) under utilitarian pressure, 2) The convention yields slowly to the moral urgency of the cause (righteousness), 3) The convention holds and rights are respected despite the consequences, and finally, 4) The convention is overridden in the face of catastrophe.

[42] Other forums, including The International Committee for Robot Arms Control refer to ALW as lethal autonomous robots or LAR. See http://icrac.net/

[43] Section 1, Chapter 1, Article 1 of the 1907 Hague IV states that armies, militia and volunteers are "To be commanded by a person responsible for his subordinates."

[44] Dabringer, Gerhard, *Ethica Themen*, 45. "The 1949 Geneva Convention requires the use of common sense while the 1977 Protocol 1 essentially defines a civilian in the negative sense as someone who is not a combatant." An inability to distinguish between a combatant and a non-combatant is an inability to properly discriminate among potential targets.

Bibliography

Alighanbari, Mehdi and How, Jonathan P. “Cooperative Task Assignment of Unmanned Aerial Vehicles in Adversarial Environments.” *2005 American Control Conference* (2005).

Arkin, Ronald C. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press, 2009.

Dabringer, Gerhard. *Ethica Themen; Ethical and Legal Aspects of Unmanned Systems Interviews*. Republik Osterreich, 2010.

Galzi D. and Shtessel Y. “UAV Formations Control Using High Order Sliding Modes.” *Proceedings of the 2006 American Control Conference*, (2006).

Government Accountability Office, “GAO/OSI-98-4 Review of USAF Investigation of Black Hawk Fratricide Incident.” 1997.

The Hague Convention 1907, “Regulations Respecting the Laws and Customs of War on Land.” http://www.icrc.org/applic/ihl/ihl.nsf/xsp/.ibmmodres/domino/OpenAttachment/applic/ihl/ihl.nsf/4D47F92DF3966A7EC12563CD002D6788/FULLTEXT/IHL-19-EN.pdf (accessed December 2013).

Jane’s Defence Online. “Aim-9X.” http://www4.janes.com/ (accessed January 2007).

Nellis Air Force Base 57th Wing. “MQ-1 Predator Fast Facts.” 2004.

Strawser, Bradley J. “Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles”, *Journal of Military Ethics Vol. 9, No. 4* (2010).

United States. Department of Defense, “Investigation Report: Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988.” Washington, D.C.: Dept. of Defense, Office of the Secretary of Defense, 1989.

United States. Department of Defense. “Unmanned Aircraft Systems Roadmap, 2005-2030.” Washington, D.C.: Dept. of Defense, Office of the Secretary of Defense, 2005.

United States. Department of Defense. “US Air Force Aircraft Accident Investigation Report: US Army UH-60 Black Hawk Helicopters 87-26000 & 88-26060 Executive Summary.” 1994.

The International Committee for Robot Arms Control, http://icrac.net/ (accessed December 2013).

Tidwell, M., “A Virtual Retinal Display for Augmenting Ambient Visual Environments.” http://www.hitl.washington.edu/publications/tidwell/index.html (accessed December 2013).

Trsek, Robert B., “Automation and Commercial Aircraft Safety.” 2003.

Turing, A. M. “Computing Machinery and Intelligence.” *Mind New Series* 59, no. 236 (October 1950): 433-460.

Vick, Alan J. and Lambeth, Benjamin S. “*RAND: Aerospace Operations in Urban Environments*.” RAND Corporation, 2002.

Walzer, Michael. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. *4th ed.* New York, NY: Basic Books, 2006.

Wikipedia, “William C. Rogers,” http://en.wikipedia.org/wiki/William_C._Rogers_III (accessed December 2013).